AI-vergreen: a multi-label Sentinel-2 training dataset for summer green and evergreen needle-leaf forest types in boreal forests biomes for remote sensing applications

Léa Enguehard[1], Birgit Heim [1], Stefan Kruse [1], Begum Demir [2], Robert Jackisch [3], Peter Christian Frandsen [1], Josias Gloy [1], Sarah Haupt [1], Laura Schild [1], Femke Van Geffen[1], Veronika Döpper [1], Ronny Hansch [4], Nicola Falco[5], & Ulrike Herzschuh [1,6]

[1] Polar Terrestrial Environmental Systems, Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Potsdam Germany; [2] Remote Sensing Image Analysis, Technische Universität Berlin, Germany; [3]Geoinformation in Environmental Planning Lab, Technische Universität Berlin, Germany; [4]Microwaves and Radar Institute, German Aerospace Center (DLR), Weßling, Germany; [5] Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory, Berkeley, CA, USA; [6] Institute of Environmental Science and Geography, University of Potsdam, Potsdam, Germany.

ALFRED-WEGENER-INSTITUT
HELMHOLTZ-ZENTRUM FÜR POLAR- UND MEERESFORSCHUNG

This presentation participates in OSPP
EGU Outstanding Student & PhD candidate Presentation contest
Sharing is encouraged

Léa Enguehard
PhD candidate
Lea.Enguehard@awi.de

I – Background/Motivation

- Boreal forests (BF), represent roughly one-third of the world's total forested area and provide critical ecosystem services including carbon stocks, climate feedback, permafrost stability, biodiversity, and economic benefits.
- They are dominated by evergreen needle leaf forests *(Pinus, Picea, Abies)* in North America and Western Siberia; and dominated by summer green needle leaf forests in Central and Eastern Siberia (*Larix*).
- Optical remote sensing (RS) applications are possible but challenging due to frequent cloud coverage, forest fires, and low illumination.
- Very few labeled datasets for RS applications focusing on BF composition and structure are available, however, such products are necessary to improve our understanding of boreal forests dynamics.

- **Here, we provide an extensive hierarchically-labeled training dataset based on Sentinel-2 image patches of boreal forests.**

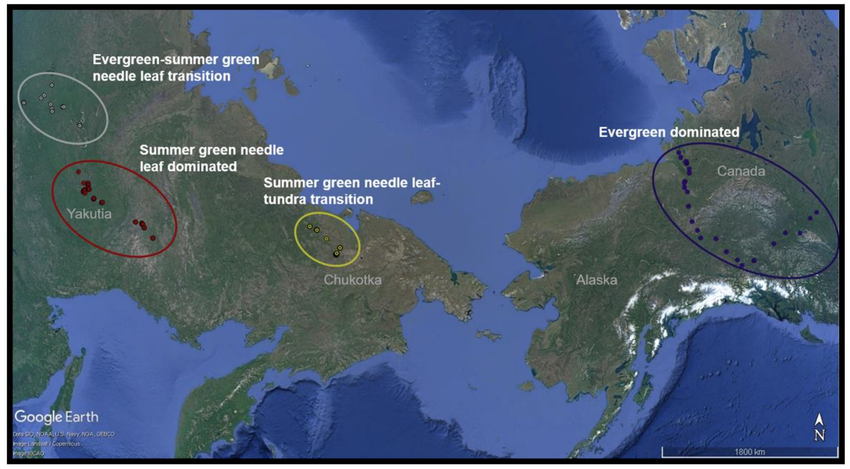

*Figure 1: Study sites visited during previous field campaigns (*RU-Land 2018, RU-Land 2021, and CA-Land 2022)

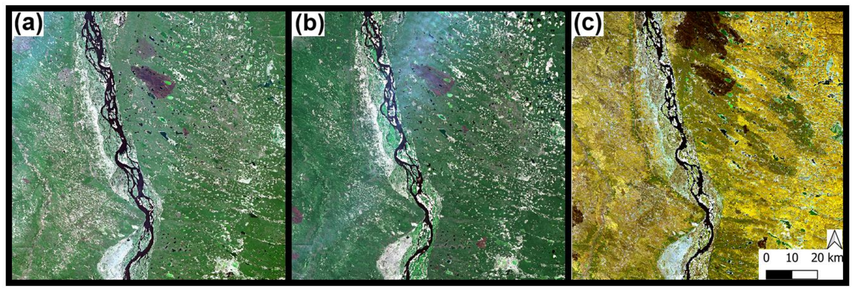

Figure 2: Example of Sentinel-2 RGB quasi-true (B4-3-2) image for the three different periods of study. The color change between (a) early (05 June 2021), (b) peak (28 July 2021), and (c) late summer (28 September 2021) helps identify specific tree species. Evergreen trees will remain green in late summer, while summer green trees will turn orange.

II – Methods

- Over 200 vegetation plots were visited during the past field campaigns of the Alfred Wegener Institute in Siberia (2018-2021) and Canada (2022), where vegetation was sampled, and described (plot extend of 60 m diameter), and UAV-borne LiDAR point clouds were collected.
- The study sites altogether cover different boreal forest types: **evergreen-summer green needle leaf transition zone** (W Yakutia), **summer green needle leaf forests** (Central Yakutia), **summer green needle leaf-tundra transition** (Chukotka), and **evergreen dominated** (Canada) (*Figure 1).*
- We gathered all cloud-free Sentinel-2 level-2 images from the corresponding year of the expedition for three periods: **early summer** (late May to late June), **peak summer** (mid-July to early August), and **late summer** (late August to September) that geographically coincide with the vegetation plots *(Figure 2).*
- For each plot and period, we selected the atmospherically corrected bands resampled to 10 m, and cropped **60 x 60 m S-2 image patches** (approx. 1000 image patches per season in the dataset as of April 2023, our work is still ongoing).
- Each 60 x 60 m plot was labeled based on tree species dominance (in-situ data) and crown cover percentage (derived from LiDAR data). K-means clustering was then applied to have more robust labels and identify outliers.

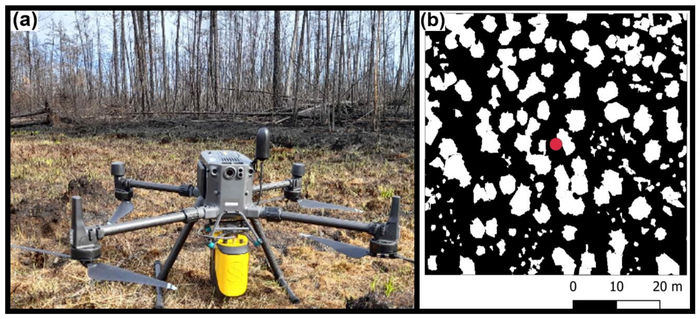

Figure 3: (a) Yellowscan mapper mounted on DJI M300 drone used to collect LiDAR pointclouds at the study sites (Photo: Robert Jackisch); (b) Example of crown cover resulting from the canopy height model (CHM) of the LidR package. Each white pixel represents tree crowns above 2 meters high, the plot center being the red dot. The crown cover in this case is 32.8%.

III – Results

We defined 9 different labels based on the tree species and 4 based on the crown cover percentage (*Figure 4*).

1. Tree species selection

- When the plot consisted of only one tree species, the label was named after this species. Eg. 60% Picea → Label = "Picea".
- When the plot consisted of two tree species with one of less than 10%, the label was named after the dominating species only. Eg. 60% Larix & Pinus 8% → Label = "Larix"
- When the plot consisted of multiple tree species with comparable coverage, the label was assigned as mixed forest.

2. Crown cover percentage

- Crown cover percentage ∈ [0 - 25[= "Very sparse"
- Crown cover percentage ∈ [25 - 50[= "Sparse"
- Crown cover percentage ∈ [50 - 75[= "Dense"
- Crown cover percentage ∈ [75-100] = "Very dense"

The labels are hierarchical up to Level 4, where only one crown cover category is assigned per label. In addition, we are currently working on a 5th Level based on the understory layer which can differ drastically between sites.

Level 0 Land type	Level 1 Forest type	Level 2 Specified forest type	Level 3 Dominating species	Level 4 Crown cover				Level 5 Understory			
				Very sparse	Sparse	Dense	Very dense	Barren/needles	Lichens	Evergreen shrubs	Summer green shrubs
Forest	Needle leaf	Summer green	*Larix*								
		Evergreen	*Picea*								
			Pinus								
		Mixed	Evergreen mixed (*Picea*, *Pinus*)								
			Evergreen - summer green mixed (*Picea*/*Pinus*, *Larix*)								
	Mixed needle leaf - Broadleaf	Summer green mixed	*Larix* - *Betula*/*Populus*								
		Evergreen – Summer green mixed	*Pinus*/*Picea* – *Betula*/*Populus*								
	Broadleaf	Summer green	*Betula*								
			Populus								
Burnt forest											

Figure 4: Hierarchical labels of the Sentinel-2 boreal forest training dataset

IV – Perspectives

- **We anticipate our Sentinel-2 training dataset to be a starting point for a significantly more extensive one** with the addition of radar satellite sensors such as Sentinel-1 and TanDEM-X, and other ground vegetation plots (new expedition expected in Alaska in summer 2023), data search in literature and repositories– e.g. NASA Arctic Boreal Vulnerability Experiment.

- Our training dataset is still in progress but will be publicly available and could be used for deep learning algorithms to identify and characterize evergreen and summer green needle-leaf trees in boreal forest regions.

The authors would like to thank all expedition members of the RU-Land 2018, RU-Land 2021, and CA-Land 2022 expeditions.
This research is supported by the BMBF AI-vergreens project and by the ERC- Glacial Legacy project (PI. Herzschuh)

POLMAR HELMHOLTZ GRADUATE SCHOOL FOR POLAR AND MARINE RESEARCH · HELMHOLTZ · TU Berlin · Universität Potsdam · DLR · BERKELEY LAB Bringing Science Solutions to the World · EGU European Geosciences Union